UNIVERSITY OF MURCIA

*International Journal of English Studies*

IJES

http://revistas.um.es/ijes

# Large-scale assessment of language proficiency: Theoretical and pedagogical reflections on the use of multiple-choice tests

IRINA ARGÜELLES ÁLVAREZ*
*EUIT de Telecomunicación - Universidad Politécnica de Madrid*

Received: 16 September 2013 / Accepted: 18 October 2013

**ABSTRACT**
The new requirement placed on students in tertiary settings in Spain to demonstrate a B1 or a B2 proficiency level of English, in accordance with the Common European Framework of Reference for Languages (CEFRL), has led most Spanish universities to develop a program of certification or accreditation of the required level. The first part of this paper aims to provide a rationale for the type of test that has been developed at the *Universidad Politécnica de Madrid* for the accreditation of a B2 level, a multiple choice version, and to describe how it was constructed and validated. Then, in the second part of the paper, the results from its application to 924 students enrolled in different degree courses at a variety of schools and faculties at the university are analyzed based on a final test version item analysis. To conclude, some theoretical as well as practical conclusions about testing grammar that affect the teaching and learning process are drawn.

**KEYWORDS**: language teaching and testing, large-scale testing, grammar tests, multiple choice tasks

**RESUMEN**
Las nuevas exigencias sobre niveles de competencia B1 y B2 en inglés según el Marco Común Europeo de Referencia para las Lenguas (MCERL) que se imponen sobre los estudiantes de grado y posgrado han llevado a la mayoría de las universidades españolas a desarrollar programas de acreditación o de certificación de estos niveles. La primera parte de este trabajo trata sobre las razones que fundamentan la elección de un tipo concreto de examen para la acreditación del nivel B2 de lengua inglesa en la Universidad Politécnica de Madrid. Se trata de un test de opción múltiple y en esta parte del trabajo se describe cómo fue diseñado y validado. En la segunda parte, se analizan los resultados de la aplicación del test a gran escala a un total de 924 estudiantes matriculados en varias escuelas y Facultades de la Universidad. Para terminar, se apuntan una serie de conclusiones teóricas y prácticas sobre la evaluación de la gramática y de qué modo influye en los procesos de enseñanza y aprendizaje.

**PALABRAS CLAVE:** enseñanza y evaluación de lenguas, evaluación a gran escala, pruebas de gramática, actividades de opción múltiple

***Address for correspondence***: Irina Argüelles Álvarez. EUIT de Telecomunicación (UPM-Campus Sur). Ctra. de Valencia, Km 7. 28760 Madrid. Tel: Email: irina@euitt.upm.es

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

## 1. INTRODUCTION

As communicative approaches to language teaching evolved (Savignon, 1977; Widdowson, 1990), communicative approaches to language testing focused on research both in relation to communicative curriculum development as well as communicative language testing (Alderson & Hughes, 1981; Lee *et al*., 1985; Nunan, 1988). The notion of "directness" had important implications for the testing of communicative performance as a "direct test" claims to measure ability directly while an "indirect test" requires the test-taker to perform more artificial tasks; interviews or role-plays to assess speaking or writing an e-mail or an essay in the case of writing are examples of direct tests (Berkoff, 1985; Connor, 1991; Cooper & Odell, 1999; Hamp-Lyons, 1995). On the contrary, grammar and vocabulary tests, usually including discrete item tasks, are typically used as indirect tests of language ability.

Provided that test users can make an easy connection between test performance and future use, direct tests usually have higher face validity[1] than indirect tests (Davies *et al*., 1999) and we agree with Nunan (1988: 117,118), that the degree to which a test appears to measure the knowledge it claims to measure should never be overestimated. Nevertheless, according to Harris (1969:21) and Oller (1979:52), this type of validity is not crucial to determine the general validity of a test. Davies (1990:23) claims that although a test should contain face validity, this must be the first one to disregard if there is any conflict with one of the other validities. But still, failure to meet face validity can eventually lead to lack of public credibility of a test as it has much to do with general acceptance.

With the awareness that the presentation of a preliminary proposal of an indirect grammar test to assess proficiency in tertiary settings could be highly unpopular in terms of face validity, at the BAAL Conference 2011 (Argüelles *et al*.) we emphasized three points in relation to its practicality:

- First, that the proposal was made for a specific context where a B2 level had to be demonstrated on the part of the students enrolling in a course of professional and academic English, with administrative aspects of the evaluation as the priority.
- Second, that it was not intended to suggest or demonstrate that an indirect test could in any case substitute direct tests of different skills; the aim was rather to present it as a practical tool where other alternatives were difficult or impossible to carry out.
- And third, that multiple choice tests are in fact a good alternative for increasing reliability when their design and development are conceived to maintain good levels of criterion-related validity. We insisted that the domain and the theoretical model presented were aimed at giving a precise response to the situation and the needs that framed our specific context.

During the debate, well-known researchers in the area of testing focused their positive comments on the adequacy of the test, which apparently showed high levels of reliability and validity, measured what it claimed to measure and therefore, could place a student above or

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

below the given B2 level of proficiency. Our concerns about the test's limitations stemming from the type of evaluation that was being carried out turned out then to have been unfounded. Today, two years later and having completed all the necessary stages of the validation process, the decision to use a grammar and vocabulary multiple-choice test seems to have been a good solution in the given context.

Two basic ideas reinforce at present the belief that a multiple choice test is a suitable answer to our needs. On the one hand, what is currently assessed under the title "grammar" is different from what was assessed as "grammar" years before. Purpura (2004: 89) defines grammatical ability as involving "the capacity to realize grammatical knowledge accurately and meaningfully in test-taking or other language-use contexts." Grammar functions at discourse level, and sociolinguistic functions can also be assessed through grammatical use as will be explained later. On the other hand, if "a prerequisite for performance is a basis of competence, no matter how minimal" (Rea, 1985: 21) grammatical knowledge or lack of it, necessarily has direct implications in communicative reception and production. Furthermore, testing is not teaching; even within highly communicative language teaching and learning contexts, language tests are still operational and their aim is to provide operational definitions of adequate language behavior. "A language test cannot therefore afford to be programmatic, to indicate what it would be useful or interesting or important or even fun to do; what it must do is to represent a decision as to what has to be done and then to do it." (Davies, 1990:15)

While the statistical validation of the test was addressed previously (Argüelles Álvarez & Pablo-Lerchundi, 2012), the statement on how the assessment of grammatical ability can be carried out, or the qualitative analysis of the final results which lead to inferences about our students' knowledge of grammar, are still lacking. Thus, in what follows, the notion of "grammar" and its role in communicative language testing will be first analysed. Then, we will summarize the rationale and validation process of the multiple choice test developed at the *Universidad Politécnica de Madrid* (UPM). The test construction was carried out in order to regulate the students' access to the subject "English for Professional and Academic Communication" for which a B2 proficiency level, in accordance with the Common European Framework of Reference for Languages (CEFRL), was established as a minimum level. Finally, the results obtained from the large-scale application of the test to 924 students in the last stage of the validation process will be analysed. A qualitative item analysis will lead us to conclusions about some of the discrete or integrated items that are not giving the expected results and, more interestingly, about eventual acquisitional sequences of grammar on the part of the students. Besides the theoretical study, these results could be useful for teachers to become more aware of some of their students' general difficulties and to introduce grammar points accordingly in their curriculum.

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

## 2. ASSESSING LANGUAGE PROFICIENCY

A few decades ago language instruction relied heavily on teaching grammar; even during the communicative era, grammar remained a key point to consider within the then "new" approach to the teaching and learning of languages (Jiménez Juliá *et al.,* 1998). With regard to its assessment, and according to Madsen (1983: 6-8), in the communicative stage, tests were concerned with evaluating communication in the second language and combined various sub-skills either orally or in writing. "Thus, the task is holistic- that is, grammar and vocabulary and overall meaning are tested simultaneously. But the scoring is quite objective". (Madsen, 1983: 7) As Madsen puts it, in this communicative approach to testing, grammatical competence also remained unquestioned as part of communicative language ability.

Today, teaching controversies revolve around the role of grammar in the language classroom and how it should be assessed accordingly. In our approach, we will adopt an interventionist position with regard to grammar instruction in the L2 classroom based on three fundamental reasons: First, because, although in the 1960s some language educators questioned the role of grammar in the L2 curriculum, most language teachers today would agree that explicit grammar instruction contributes to students' linguistic development; second, because, although research gives credit to some of the non-interventionist claims (Purpura, 2004: 32-34), empirical research in Second Language Acquisition (SLA) confirms that the instruction of L2 grammar is effective (Doughty & Williams, 1998) and third, because in our experience, most students welcome explicit grammar explanations when particular points need clarification. Regarding its assessment, in many testing contexts today, knowledge of grammar is inferred from the ability to select a correct option among several or the ability to use it correctly while reading, or speaking but "[...] there is a glaring lack of information available on how the assessment of grammatical ability might be carried out, and how the choices we make in the assessment of grammatical ability might influence the inferences we make about our students' knowledge of grammar, the decisions we make on their behalf and their ultimate development." (Purpura, 2004: 4).

According to Ellis (2001), findings from SLA research are helpful to language testers to further reflect on the design and development of grammar tests (see Long, 2011: 378-381 for a review of research findings during the last 40 years). Notions such as "implicit knowledge" (intuitive and rapidly processed) and "explicit knowledge" of grammar in the form of metalanguage or analysed knowledge, raise crucial questions concerning the type of knowledge testers want to test and how they do it. A case for testing a combination of both implicit and explicit knowledge in the form of analysed knowledge could be, according to Ellis (2001: 252), a population of learners planning to enroll in an academic program. Within this context, he suggests that pressurizing students under a time constraint while performing a discrete-item grammar test will force them to draw on their implicit knowledge.

Another principal finding of SLA research is the order of acquisition of grammatical

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

structures as there is convincing evidence that learners progress through an order (Ortega, 2011) and therefore, "[...] grammatical structures are not equivalent in difficulty, if difficulty is equated with the order in which structures are acquired." (Ellis, 2001: 254). As a result of this finding, the notion that structures are acquired in a fixed sequence should also have some relevance to the assessment of grammatical ability.

Based on the former and other studies he presents, regarding the effects of grammar teaching and learning in language assessment, Purpura (2004: 45, 46) highlights the need on the part of test developers to inform readers about the test specifications and to provide technical information on the quality of the assessment. In line with his recommendation, in what follows, detailed information about the development of the test, test specifications, pre-pilot and pilot stages and final large-scale test application, will be analysed. Later, in the "results" and "discussion" sections we will address aspects of grammatical knowledge that were assessed by means of different tasks and draw theoretical conclusions about the students' grammatical ability which will influence subsequent variations of the original test. Furthermore, according to our view of testing as completely integrated with the teaching-learning process, we will suggest possible instructional solutions to deal with some of the most general problems detected that regard knowledge of the L2.

## 3. TEST DESIGN AND VALIDATION

### 3.1. Test layout and description

The degree changes affecting engineering studies at *Universidad Politécnica de Madrid* (UPM), led four years ago to the implementation of a new compulsory subject across all its schools and faculties, English for Professional and Academic Communication. The University also established a B2 proficiency level, in accordance with the Common European Framework of Reference for Languages (CEFRL), as a minimum level for students to enroll in the subject. In the context described, a proficiency test was developed for our "real world purpose" (Davies *et al*., 1999: 154) mainly pursuing an administrative aim to regulate the students' access to the subject, by placing them above or below the B2 level. Our specific context, the number of students who must certify a B2 level and the heterogeneous background of the more than sixty teachers in the Department of Linguistics, lead us to opt for an automatic correction test made up of multiple choice- type questions. For the same practical reasons of application, the test does not include a listening comprehension section or an interview and its limitations regarding students' results were studied and assumed in the planning stage (Argüelles Álvarez & Pablo-Lerchundi, 2012). The test was designed to check that the students have the required minimum previous knowledge and not to place them within a scale; the students do not receive feedback or information about their proficiency level within the standards established by the CEFRL, they only receive a message saying

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

whether or not they have reached the threshold of the level required to enroll in the subject, "English for Academic and Professional Communication."

According to Davies (1990: 13) there is no agreement in language test construction as to the importance of the language system or the use of the language system. Therefore, in order for the multiple choice test to reflect both a psychometric-structuralist view of language test construction (Spolsky, 1977 in Davies, 1990) as well as a psycholinguistic-sociolinguistic view, the test was divided into two parts. The first part, consisting of 65 individual items followed by the four options a, b, c and d, evaluates aspects of grammar concentrating mainly on the language system, although pragmatic meaningfulness or acceptability are also contemplated herein. Language functions such as showing agreement or disagreement and language notions such as point of time (*since*) versus period of time (*for*) are also assessed by means of grammatical and lexical forms. The assessment of grammatical ability is therefore strongly based on functional-notional categories which, according to Purpura (2004: 20), have substantial impact on L2 syllabus design and are recognized for shifting the emphasis from a syntactocentric perspective to a communication-based one.

The second part of the test deals more with the use of language system and consists of three texts, two assessing grammar and vocabulary in text as input and the third, reading-comprehension skills, each assessed out of 10 or 15 points, for a total of 35 points. Regarding the first two texts the task type is a rational multiple choice cloze (Madsen, 1983:23) where the test designer decides which words to omit and adds alternate responses from which the students must choose the correct one to fill in the gaps in the text. According to Bensoussan and Ramraz (1984), the multiple-choice version of a rational cloze (the fill-in test) lets the test designers focus on text micro or macro level thus directing the responses to suit their specific needs. Students can therefore be forced to relate the information with "extralinguistic context" or to predict information for the gap including a discourse level that takes into account cohesion and coherence aspects.

The last text with comprehension activities (questions followed by four possible answers) forces students to establish the necessary relationships among grammatical form, grammatical meaning and pragmatic meaning. Researchers who have investigated students of English as a foreign language taking reading tests, note that they "use a combination of prior knowledge, analysis of the text and accompanying questions, and test taking skills" (Allastir, 1992:101), calling upon the so called multi-componential language ability in Bachman and Palmer (1982). The multi-componential model viewed language ability as an interaction of language knowledge with non-linguistic components such as topical knowledge, personal characteristics and strategic competence.

In order for the test to present the same standard of difficulty in its different versions, a sufficient number of items were stored in a resource repository, a bank of items and a bank of short texts, so that different alternatives were offered to assess the same aspects of language. At that stage, we counted therefore, on an adequate pool of items, an inventory of the abilities

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

each item purported to measure and a construct of the abilities tested which would permit the score interpretation (Davies *et al*., 1999: 93).

Although the test was designed for the b-learning platform Moodle, and tried out using the platform at the pre-pilot and pilot stages, lack of technological support for large-scale testing led us to print different versions and deliver these on paper to the students at the final validation stage. The items were stored in the platform, organized in blocks relating to aspects of language (grammar/use of language, vocabulary, and reading) and in sub-topics (grammatical, functional or topical/vocabulary-related) within these aspects. For each of the versions, the program would choose a predetermined number of activities from each of the blocks. That is to say, the program would make a semi-random selection of items, limited by the condition that it chose one item from each of the groups.

### 3.2. Validation process

As described by Davies (1990: 12-13), prepilot, pilot (or pre-test) and final validation stages are procedural stages where techniques of item analysis and descriptive statistics are employed and are partly dependent on the success of a previous first stage: planning. Regretfully, and according to the same author, information on the preliminary stage is usually limited and the process of item planning and writing is hardly ever described in any detail although the planning stage is critical for language sampling and test validation. The planning stage opens our description of the validation process below.

As a proficiency test, our test does not exhibit any control over previous learning but establishes generalizations from a basis of typical syllabuses to make it more directly connected to what it aims to assess. In our case, the activities included in the test were adapted into a multiple choice format with four options from a corpus of texts and tasks selected from general English course books correlated to the CEFRL, covering from B2 towards C1 levels. From the corpus, the core vocabulary, grammatical structures and functions, and the difficulty of the texts for the level were established. For the first part of the test, 65 items would stand alone under a general instruction ("Please select the best answer from the options provided") and a stem (a phrase, a sentence a question or short dialogue to be completed). The second part of the test, presents the stimulus material in the form of three texts. The first two texts propose cloze-type tasks and reading comprehension questions follow the third text.

In the first revision, once the activities had been adapted, native-speaker teachers of English with experience in testing checked that the items were clear and unambiguous for content validity[2]. At the same time, they made sure that only one of the options given could be correct or clearly more suitable to the given context. Apart from the correct option, the other three options were adapted to the following scheme: one answer seemed very likely although it could not be possible and the other two were not possible. One of these last two represents, when possible, common error tendencies detected among Spanish learners of

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

English as a foreign language usually as a result of interference from the native language or other factors, observed during years of prior teaching experience. Up to 35% of the items or the options given as possible solutions were discarded or modified at this stage during a rather time-consuming process.

During the pre-pilot stage, the test was tried out in different, less formal to more formal trials on subjects who represented the target test population at the UPM School of Telecommunications. As explained in Davies *et al*. (1999: 150), the purpose of pretesting is to identify problems in any aspect of the testing procedure. The final pilot test consisted of 100 items selected from among approximately 1,000 validated items which made up the initial bank. In this case, for research reasons, the 100 items selected were the same for all the students taking the pilot test whereas, at the final validation stage, the selection of items was made randomly by the program for each version of the exam (two in the first large-scale test). The test was tried out in an experimental situation with first-year students in the UPM School of Telecommunications (Campus Sur) as representative of the sort of students who would take the test in the future. Then, results were studied to reach conclusions concerning the test reliability and validity. As these data were published in 2012 (Argüelles Álvarez & Pablo-Lerchundi) we will only summarize here the final conclusions and briefly comment on these two fundamental concepts.

At the pilot stage, a total of 240 incoming students at School of Telecommunications took the test and a total of 214 tests were taken into consideration for the statistical analysis. As the test layout is in two well differentiated parts—the first part focused on discrete items while the second is based on text input— a split-half method was used to estimate the reliability of the test, or "The actual level of agreement between the results of one test with itself or with another test" (Davis *et al*., 1999: 168). The Pearson correlation coefficient between the scores obtained in part one and those in part two of the test was analyzed and the result was 0.81, which is statistically significant (see Table 1). This means 81% concordance between the two parts of the test.

| | Part two (use of language, vocabulary and reading) |
|---|---|
| Part one (grammar) | 0.813** |

**p< 0.01

**Table 1.** Pearson Correlation between part one and part two of the English Proficiency Test

As regards validity or "the extent to which it [the test] succeeds in providing an accurate concrete representation of an abstract concept (for example proficiency, achievement, aptitude)" (Davies *et al*., 1999:221), a statistically representative sample of 31 students from the 214 students who completed the multiple choice test was selected for an interview. The examiner, unaware of the previous results of these 31 students, held a personal ten-minute interview with each of them. The correlation between the results in the interview

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

and the test provided a validity coefficient of 0.83, significant at $p< 0.01$.

| | English Proficiency Test |
|---|---|
| Oral Interview | 0.825** |

**p< 0.01

**Table 2.** Pearson Correlation between Oral interview and English Proficiency Test

At the final validation stage, two random versions of the test consisting of 67 discrete items with phrase or sentence length input (part 1) and 33 items in extended discourse (part 2) were completed. The test took place across the university and was taken by 924 students enrolled on different degree courses at all its schools and faculties. As indicated previously, lack of technological support in the different schools where the test took place, led us to print the two versions and deliver these on paper to the students together with a computerized answer sheet at the final validation stage. The students were given 50 minutes to complete the test.

Based on absolute standards, previous analysis (Argüelles Álvarez & Pablo-Lerchundi, 2012: 16) and possible classification errors (Bachman, 1990: 75) a cut-off score of 68 points was established as the line between mastery and non-mastery on this occasion. In order to avoid false negative classification errors, once the tests had been corrected, the cut-score was lowered to 65 after comparing the percentages of students who had passed with those obtained from students who had been considered "qualified" in previous pre-pilot and pilot stages.

## 4. RESULTS

"Strictly speaking, criterion-referenced tests are only concerned with whether candidates have reached a given point rather than with how far above or below the criterion they may be." (Davies *et al*., 1999: 38) In other words, criterion referencing is a way of determining, for a given situation, how much enough is. Items that are too easy (with an index close to 100%) or too difficult (with an index close to 0%) do not contribute to a test's discriminability, especially important in norm-reference tests, and are therefore normally discarded in the latest type of tests. But the dichotomy easy/difficult can be useful in criterion-referenced measurement for clarifying teaching objectives or for research purposes, as is the case herein. The degree of difficulty of a test item, calculated on the basis of a group test performance, can eventually lead us to conclusions about the degree of difficulty of the trait under test.

Although for multiple-choice tests, the average item difficulty index is set higher to compensate possible guessing strategies, generally speaking, standardised tests aim at a range of 30% to 70% spread of difficulty, averaging out at approximately 50% (Davies *et al*., 1999: 95, 96). These percentages are the starting point for our analysis of the results, which in the

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

present study will be mostly centred on the analysis of the two extremes of difficulty.

From our data, the following are the results obtained after the application of the test regarding the difficulty of the items in the two versions of the tests (Figure 1 and Figure 2):

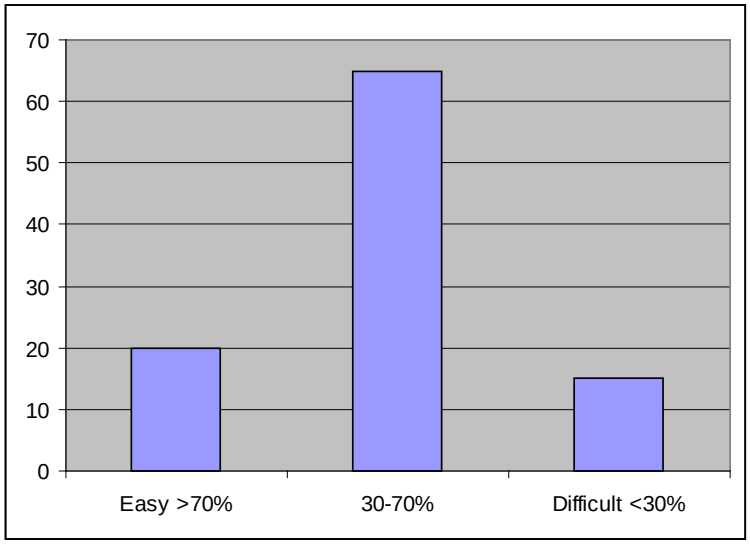


**Figure 1.** Final results of item difficulty in Version 1 of the test

As regards Version 1 of the test, 20 items are considered in the category of "easy" having been answered correctly by more than 70% of the students. At the other extreme, 15 items fall in the category of "difficult" items as less than 30% of the students answered these questions correctly. 65 items are therefore within the most desirable range of 30% to 70%, and all the items in the test average out at 51.36%.

Focusing on the difficult items in Version 1 of the test, 9 items out of the total number of 15 in this category correspond to answers to contextualized items, those that are in text in the second part of the test, evenly distributed among the three of them. From the 20 items in the category of "easy", only three are found in the second part of the test. These results are summarized in Table 3:

| | Correct answers <30% | Correct answers >70% |
|---|---|---|
| Discrete items with length of phrase or sentence (67 items) | 6 | 17 |
| Text 1 (10 items)<br>Multiple-choice cloze<br>Selected response | 4 | -- |
| Text 2 (15 items)<br>Multiple-choice cloze<br>Selected response | 3 | 3 |
| Text 3 (8 items)<br>Reading comprehension questions<br>Selected response | 2 | -- |

**Table 3.** Version 1 difficult/easy items in the second part of the test (input as text)

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

In Version 2 of the test, 27 items fall into the category of "easy". At the other extreme, 23 items fall in the category of "difficult" with less than 30% correct answers registered. 50% of the items are therefore within the most desirable range of 30% to 70%, and all the items average out at 51.25%.

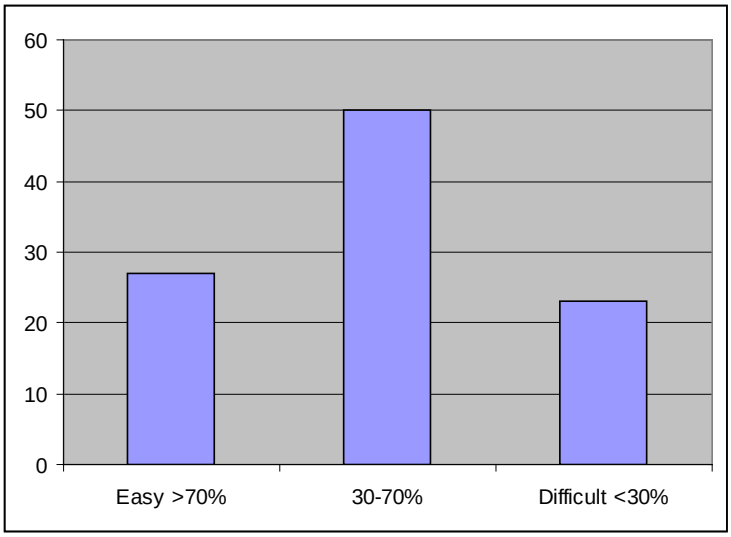


**Figure 2.** Final results of item difficulty in Version 2 of the test.

From the 27 items in the category of "easy", only two are found in the second part of the test. With regard to the difficult items in Version 2 of the test, 16 items out of the total number of 23 in this category correspond to answers to contextualized items, those that are in text in the second part of the test. Eleven items out of these 16 are located in the same multiple-choice cloze activity with text as input. These results are summarized in Table 4:

| | Correct answers <30% | Correct answers >70% |
|---|---|---|
| Discrete items with length of phrase or sentence (67 items) | 7 | 25 |
| Text 1 (10 items)<br>Multiple-choice cloze<br>Selected response | 3 | 2 |
| Text 2 (15 items)<br>Multiple-choice cloze<br>Selected response | 11 | -- |
| Text 3 (8 items)<br>Reading comprehension questions<br>Selected response | 2 | -- |

**Table 4.** Version 2 difficult/easy items in the second part of the test (input as text)

## 5. ITEM CLASSIFICATION

In order to reach preliminary conclusions from the results obtained, we are mainly concerned with the classification easy/difficult of the discrete items that make up the first part of the test.

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

Firstly, we will review the items that were classified as easy, and secondly we will address the difficult ones. In both cases, as it was previously stated, we are dealing with items that do not contribute to the test's discriminability but could be suitable examples from where to infer common areas where students, generally speaking, are more proficient (easy items) or less proficient (difficult items).

### 5.1. Easy items (>70%)

Among the grammar, functions or notions addressed in the 13 items that are answered correctly on the part of the test takers in the range of >70%, many address temporal and aspectual meanings (anaphoric time, duration or frequency) as well as notions related to time and temporality such as grammatical tense and aspect as shown below (Examples 10-13):

Example 1: (Item 11 Version 2) [...] already*[...]
Example 2: (Item 7 Version 2) [...] during*[...]
Example 3: (Item 35 Version 1) [...] usually get up*[...]
Example 4: (Item 36 Version 1) [...] have ever known*[...]

Others worth mentioning are grammatical knowledge and use of subordinating conjunctions as in Example 14:

Example 5: (Item 10 Version 1) [...] unless* you press the bell.

Finally, with regard to modal verbs, those indicating "impossibility" are classified in this range, while "certainty" falls in the category from 30% to 70%. See examples 15 and 16 below:

Example 6: (Item 32 Version 1) You _____ go wrong if you follow the instructions. *Impossible*. Options: a) might, b) must, c) could, d) can't* (80.44% correct answers)
Example 7: (Item 33 Version 1) He _____ have taken the money. *Certain*. Options: a) may, b) must*, c) could, d) can't (51.56% correct answers).

### 5.2. Difficult items (<30%)

#### *5.2.1. Little* vs. *a little*

Item 13 addresses grammatical form and meaning of *few*, *a few*, *little*, *a little,* those being the four options to fill the gap in the sentences provided. In both versions of the test, students fail to select *a little** as the correct option, as seen in the examples below:

Example 8: (Item 13 Version 1) I only had _____ money left and I decided to spend it on a gift for my grandmother.

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

Students fail to identify *a little* as the correct answer while they are able to see *little** as correct in its context according to the results in the item that follows number 13:

> Example 9: (Item 14 Version 1) There is very _____ hope now that war can be avoided. (Results in the range of 30%-40%)

### *5.2.2. I'd rather*

Item 17 is presented in the form of an adjacency pair to assess grammatical form in the context of the adverb *rather* used as "more readily or willingly" (Merriam-Webster)

> Example 10: (Item 17 Version 2) Q: Shall I stay here? A: I'd rather _____ with us. Options: a) you come, b) you to come, c) you came*, d) you would come

### *5.2.3. So and neither*

Designed to test grammatical form and meaning (cohesive-ellipsis), item 26 also seeks the correct function to express "in a similar manner or way" where students fail to identify *so* as the correct answer when the sentence provided as input is positive:

> Example 11: (Item 26 Version 2) My father works at home and _____ does my mother. Options: a) so*, b) neither, c) either, d) same.

On the contrary, students answer within the range of 55%-70% when the sentence given as input is negative and form among the same options as in the previous item:

> Example 12: (Item 27 Version 1). I haven't tried speed dating and _____ have my friends. Options: a) so, b) neither*, c) either, d) same.

### *5.2.4. Syntactic accuracy (word order)*

Item 22 was designed to test grammatical form of the genitives and word order but while in Version 2 of the test students give the correct answer at a 48.00%, in Version 1 this item turns out to be the most difficult in the version: only 8,13% of the test takers answered the item in Version 1 correctly.

> Example 13: (Item 22 Version 2) This isn't my book. It's _____ . Options: a) my sister's*, b) my sister, c) of my sister, d) of my sister's
> Example 14: (Item 22 Version 1) It was _____ to go fishing. Options: a) a good idea of Peter's, b) a good idea of Peter, c) Peter's good idea*, d) Peter's a good idea

From the classification easy/difficult, we have provided a first descriptive approximation to those areas where most students in our context are more proficient and those areas where they may find eventual difficulties. The different subsections presented above define the general areas under scrutiny which have been deduced and formulated from the

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

results obtained in a number of discrete items. Then, some of these items have illustrated, in the examples, the general aspects that have resulted to be easier (5.1) or more difficult (5.2) for the students. In the discussion section, we will analyze the former and some of the items that seem to be working differently in the texts as inferred from the results shown previously in Table 4. The examination of the content of individual items within the most desirable difficulty range of 30% to 70% and their discrimination values is left for future research.

## 6. ANALYSIS AND DISCUSSION

The first thing we notice while observing the results from Version 2 of the test is that the second text with filling-in activities is not giving the results expected (11 items out of 15 are classified in the range of <30%, or "difficult"). We will explain the high level of difficulty of the second text in Version 2 of the test as derived either from its content or the language, based on the theory of pragmatic expectancy grammar developed by Oller (1979). According to this theory, the student must be able to process sequences of language and to understand the pragmatic interrelationship of linguistic and extralinguistic contexts. If difficulty originates in the content it could be explained by the failure of students to map the topic onto their own personal experience. If it comes from the language, it would mean an eventual failure of students to map the vocabulary onto their previous L2 learning experiences. As the topic of the text is personality / psychological response, *Dealing with regret*, our conclusion is that the problem is more with the language used rather than with the topic. Apart from personality adjectives, the gaps address other more specific idioms (*It's no use crying over split milk*) or word collocations (*to lose one's temper*) related to behaviour.

Although we recognize here a problem of test construction, as the text is heavily based on responses about a very specific topic, it is worth mentioning that, while it is certainly difficult to preview such content or language problems, the test random selection of items, including this text, has still produced, on the whole, the desirable levels of items averaging 51.25%. The general level of difficulty of the test remains the same as that in Version 1 where similar problems were not detected.

Looking at the first part of the test, discrete items, patterns could be inferred in the case of groups of words with similar meaning which could indicate that there are meanings or functions that are acquired earlier than their pair or counterparts. See, for example, the case with *little*/ *few* and *a little*/ *a few*. In general, students are able to use the first pair better than the second and only 17.33% of the students chose the correct answer, *a little*, in Example 1. The same happens with the pair *so* / *neither*; here, students fail to use the particle in the positive sentence whereas they use it correctly when the sentence is negative. Failure to give the correct answer in the affirmative sentence (18.45% in Version 1 of the test and 17.11% in Version 2) but not in the negative (57.33% and 67.46% respectively) can look strange at first

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

as it is usually the non-marked structure (Greenberg, 1966) which is commonly supposed to be more readily learnt by the L2 student (Rutherford, 1982). Furthermore, if we look at the correct use of *so* when paired or grouped with other particles for emphasis (*so*/ *such*/ *such a*), results reach 80.44% in Version 1 and 57.74% in Version 2 of the test. Items 3, 4 and 5 seek to identify grammatical knowledge and use of this and other particles for emphasis. In general, students have been able to identify the correct answer in: *so good looking* or *so much fun*. These results, point to other possible factors intervening in the acquisitional sequence apart from markedness, as suggested by Bardovi-Harling (2006) or Haspelmath (2006), which exceed the scope and limits of our work but are worth studying in depth. Awareness on the part of teachers of these and other acquisitional sequences and the reasons that justify such sequences would permit the prediction of possible learning problems. Thus, a more realistic lesson plan could eventually be developed if these aspects were to be treated within the program or syllabus.

To finish the review of difficult discrete items, learning sequences can also be inferred from the last example presented in the results section where accuracy in word order depends very much on the syntactic difficulty of the structure. Only 8.13% of the test takers chose the correct answer when the genitive is part of a complex sentence as is the case with the extraposition of a clausal subject: *It was Tom's good idea to go swimming*, while 48.00% of the students answer correctly when the genitive is found in a simple sentence: *This isn't my book. It's my sister's.*

Moving on to classroom teaching, scores from the test have been useful not only in deducing grammatical accuracy, but also in making inferences about the underlying acquisitional development of the L2 learners. The results from the item analysis can lead to classroom concerns such as what grammar to teach and making use of grammatical sequencing criteria (Canale & Swain, 1980: 32). Positive backwash (or washback) derives from such concerns if teaching materials and methods progressively integrate activities and tasks related to the grammar areas that have proved to be more problematic for students (Alderson & Hamp-Lyons, 1996). Then, any change in the syllabus and materials should ultimately have an effect on subsequent test item construction. According to Davies (1985: 8), innovative proficiency tests ideally produce a syllabus change that transforms a proficiency test into an achievement test and thereby a new proficiency test must be constructed which favours further the development of new ideas for language teaching and learning. This two-way relationship between language teaching and language testing implies the design of new materials and the incorporation of innovative methods that pedagogical grammars or the latest text books address explicitly.

Usually, such pedagogical grammars deal with these more problematic aspects of language adding novelties in the instructional techniques derived from research in innovation. These techniques can ease the inclusion of grammatical aspects such as the ones mentioned, in a communicative approach to the teaching and learning of languages. According to Purpura

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

(2004: 40, 41), four types of instructional techniques focus the latest research on teaching grammar:

- Form-based or rule-based techniques involve implicit, inductive as well as explicit deductive grammar teaching and involve consciousness-raising activities.
- Input-based techniques are based on the use of input for grammar instruction, where learners are asked to relate grammatical form and meaning.
- Feedback-based techniques rely on negative evidence of grammar performance when a generalization does not hold.
- Practice-based techniques involve input processing instruction and output practice.

## 7. FINAL CONCLUSIONS

Besides the formal study presented herein, grounded on the design and validation of a multiple choice grammar test, our argument in this paper has been that results can be practical for teachers. Results from the test highlight some of the students' general difficulties which might help instructors introduce grammar points accordingly in their curriculum. Findings about eventual developmental orders in processing, such as the ones mentioned in SLA research, can be amplified, based on further test applications. This will necessarily have a positive effect in grammar instruction within communicative programs and in subsequent testing processes as a natural consequence. The view of language testing as completely integrated in the teaching-learning process is not new in our case (Argüelles Álvarez, 2012:130) and the study presented herein is just another example that illustrates their inextricable relation.

Although we have only drafted here possible areas where sequential learning on the part of students could be studied at the intermediate developmental stage, it seems that there is much work to do to explain these and other additional patterns in L2 development that affect the teaching of grammar. With regard to further research in this area, corpus linguistics has much to offer test developers regarding test content and item design (Biber *et al*., 2004) and can also shed light on some unsolved questions regarding possible factors intervening in the L2 learning acquisitional sequences.

Returning to face validity, it is still believed that these types of tests will continue to be seen, at least for the moment, as old-fashioned and definitely non-connected with current teaching contexts. While developing this large-scale test program, full understanding has been reached on the part of the researcher, that it is complex to make explicit the exact relation of the test tasks designed with a communicative approach to the teaching of languages. Nevertheless, as a final conclusion, it is difficult to avoid insisting that this sort of testing has, in reality, been useful and adequate for our purposes, and raises new questions with regard to large-scale testing and its relation with language teaching.

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

## NOTES

1. With regard to validity, in any context of evaluation, tests need to fulfill four criteria of validity (Ferris & Hedgcock, 1998: 230- 232):
   - Face validity: implies that both students and teachers understand the instrument as adequate for what it is intended to assess.
   - Criterion validity (*concurrent validity* in Davies (1990: 24), Jacobs *et al.* (1981:74) and Oller (1979:51)): If the instrument is valid, it will produce similar results to those obtained from the application of another test already validated and administered under similar conditions.
   - Content validity: It is related to the efficiency of the method to force the students to demonstrate their command in the specific area tested.
   - Construct validity: is an indication of how much the test evaluates the skill or ability which it claims to measure.

2. I want to thank the teachers at the University of Leeds Language Centre who collaborated at the revision stage and especially Ms. Stazicker for her encouraging comments and feedback.

## REFERENCES

Alderson, J. C. & Hughes, A. (Eds.) (1981). *Issues in language testing, ELT documents 111*. London: The British Council.

Alderson, J. C. & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13,* 280-297.

Allastir, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing, 9*(2), 101-119.

Argüelles Álvarez, I. (2012). El resumen y su evaluación: Aspectos teóricos y pedagógicos en el contexto de lenguas extranjeras. *ELIA: Estudios de Lingüística Inglesa Aplicada, 12,* 115-152.

Argüelles Álvarez, I., Pablo-Lerchundi, I., Herradón Díez, R. & Baños Expósito, J.M. (2011). Large-scale Testing of Proficiency in English: Back to Multiple Choice? *Proceedings of the BAAL Conference* (pp. 13-16). University of the West of England..

Argüelles Álvarez, I. & Pablo-Lerchundi, I. (2012). …And back to multiple choice! Large-scale testing of proficiency in English: an experience. *ODISEA, Revista de Estudios Ingleses, 13,* 9-18.

Bachman, L. F. & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bardovi-Harling, K. (1987). Markedness and salience in second language acquisition. *Language Learning*, 37 (3), 385-407. Article first published online: 27 OCT 2006 DOI: 10.1111/j.1467-1770.1987.tb00577.x

Bensoussan, M. & Ramraz, R. (1984). Testing EFL reading comprehension Using a multiple-choice rational cloze. *The Modern Language Journal, 68*(3), 230-239.

Berkoff, N. A. (1985). Testing oral proficiency: A new approach. In Y. P. Lee, A. Fok, R. Lord, & G.Low (Eds.), *New Directions in Language Testing* (pp. 93-100). Oxford: Pergamon.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Hetl, M., Clark, V., Cortes, V., Csomay, E. & Urzua, A. (2004). *Representing Language Use in the University: Analysis Of The TOEFL 2000 Spoken And Written Academic Language Corpus*. TOEFL Monograph, 25, Princeton: Educational Testing Service.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Connor, U. (1991). Linguistic / rhetorical measures for evaluating ESL writing. In L. Hamp-Lyons, (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 215-225). Norwood, New Yersey: Ablex Publishing Corporation.

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131

Cooper, C. R. & Odell, L. (Eds.) (1999). *Evaluating Writing: The Role of Teacher's Knowledge about Text, Learning, and Culture*. Urbana, Illinois: NCTE.

Davies, A. (1985). Follow the leader: Is that what language tests do? In Y. P. Lee, A. Fok, R. Lord, & G. Low (Eds.), *New Directions in Language Testing* (pp. 3-13). Oxford: Pergamon.

Davies, A. (1990). *Principles of Language Testing*. Oxford: Basil Blackwell.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Doughty, C. & Williams, J. (Eds.). (1998) *Focus on Form in Classroom Second Language Acquisition*. Cambridge: Cambridge University Press.

Ellis, R. (2001). Some thoughts on testing grammar: an SLA perspective. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with Uncertainty: Essays in Honour of Alan Davies* (pp. 251-263). Cambridge: Cambridge University Press.

Ferris, D. & Hedgcock, J. S. (1998). *Teaching ESL Composition: Purpose, Process and Practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Greenberg, J. H. (1966). *Language Universals*. The Hague: Mouton.

Hamp-Lyons, L. (1995). Rating nonnative writers. The trouble with holistic scoring. *TESOL Quarterly, 29,* 759-762.

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw Hill.

Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics, 42*(1), 25-70.

Jacobs, H.L., S.A. Zinkgraf, D.R. Wormuth, V. F. Hartfiel & J.B. Hugey (1981). *Testing ESL composition: A practical approach.* Rowley MA: Newbury House.

Jiménez Juliá, T., Losada Aldrey, M. C. & Márquez Caneda, J.F. (Eds.). (1998). Español como lengua extranjera: enfoque comunicativo y gramática. *Actas del IX Congreso Internacional de ASELE*. Santiago de Compostela: Centro Virtual Cervantes.

Lee, Y.P., Fok, A., Lord, R. & Low, G. (Eds.). (1985). *New Directions in Language Testing*. Oxford: Pergamon.

Long, M. H. (2011). Methodological principles for language teaching. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 373-394). Oxford: Blackwell Publishing.

Madsen, H. S. (1983). *Techniques in Testing*. Oxford: Oxford University Press.

Merriam-Webster on-line at http://www.merriam-webster.com/dictionary [last consulted on 22/08/2013]

Nunan, D. (1988). *The Learner-centred Curriculum*. Cambridge: Cambridge University Press.

Oller, J. W. Jr. (1979). *Language Tests at School*. London: Longman.

Ortega, L. (2011). Sequences and processes in language learning. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 81-105). Oxford: Blackwell Publishing.

Purpura, J. E. (2004). *Assessing Grammar*. Cambridge: Cambridge University Press.

Rea, P. M. (1985). Language testing and the communicative language teaching curriculum. In Y. P. Lee, A. Fok, R. Lord, & G.Low (Eds.), *New Directions in Language Testing* (pp. 15-32). Oxford: Pergamon.

Rutherford W. E. (1982). Markedness in second language acquisition. *Language Learning, 32*(1), 85-108.

Savignon, S. (1977). *Communicative Competence: Theory and Classroom Practice*. New York: McGraw Hill.

Spolsky, B. (1977). Language testing: Art or science? In G. Nickel (Ed.) *Proceedings of the Fourth International Congress of Applied Linguistics*, 3. Stuttgart: Hochschulverlag.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.

 
Print ISSN: 1578-7044; Online ISSN: 1989-6131